Long-Term Preservation of the Finnish Web Archive

Petteri Veikkolainen and Lassi Lager
(The National Library of Finland)
with
Juha Lehtonen (CSC - IT Center for Science)
Kaisa Kaunonen (National Library of Finland)

IIPC General Assembly, Reykjavik, 12 April, 2016

Photo: Kati Winterhalter

The National Library of Finland Collects and Preserves…

- Printed legal deposits (1707-)
- Digitized collections
- **Electronic legal deposits**
- **Finnish Web**

The Constitution of Finland (*Section 20)*

- *Nature and its biodiversity, the environment and **the national heritage** are the responsibility of everyone.*

Law on collecting and preserving cultural materials (1433/2007)

- *The National Library has the duty to collect and preserve **representative** and **diverse** range of **online material** on servers in Finland, or elsewhere if the material is targeted for the Finnish public*

Most of the photos: Renovation and preservation of the National Library of Finland 2013-2015.

The Finnish Web Archive

- **Annual Crawls**
  .fi and .ax domains + selectively other Finnish domains
- **Daily/weekly Crawls**
  Finnish news sites, e-journals etc
- **Thematic harvesting**
  Important, unexcepted, or just generally interesting national and international events and phenomena.

- **Separately collected, but not in Web Archive:**
  Institutional repositories (OAI/PMH),
  deposited e-publications

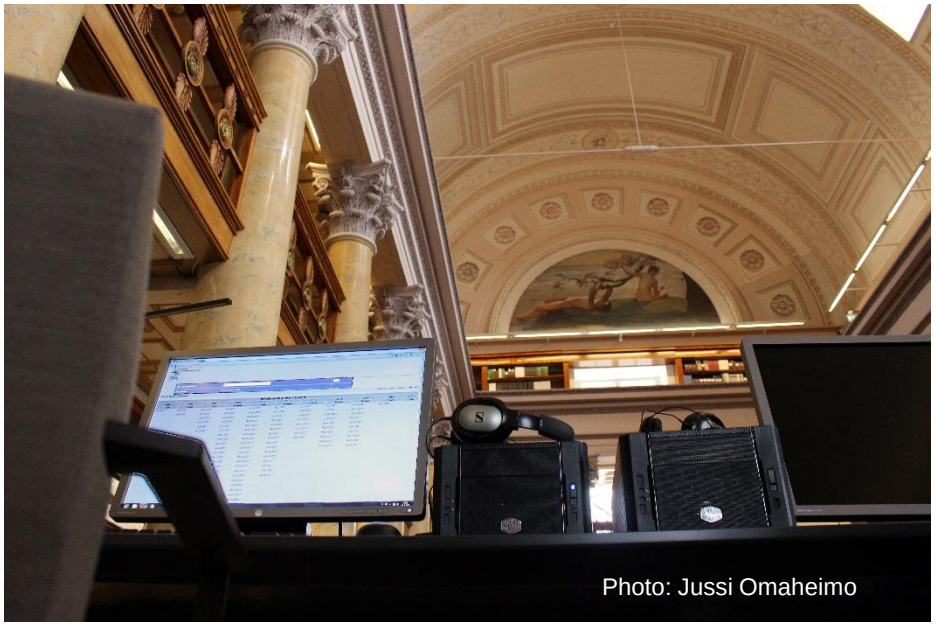


Photo: Jussi Omaheimo

Some examples of thematic harvesting

- World Design Capital Helsinki, 2012
- Railways in Finland 150 years, 2012
- Finnish presidential election, 2006 and 2012
- Summer festivals, 2011
- Turku 2011 European Capital of Culture, 2010 and 2011
- Eruption of Eyjafjallajökull volcano in Iceland, 2010
- Websites of Finnish municipalities, 2013
- European Elections, 2009 and 2014
- Finnish parliamentary elections, 2007, 2011 and 2015
- “European Refugee Crisis” 2015

Polling station at Herttoniemi elementary school (1960).

The National Digital Library (NDL)

Why?

To ensure, that electronic materials of Finnish culture and science are

- easily accessed
- **securely preserved well into the future.**
- managed with a high standard

How?

- Common user interface FINNA for libraries, archives and museums
- **A digital preservation solution for digital cultural heritage**
- Promotion of interoperability of information, processes and IT systems in Finnish memory organizations

http://wwwkdkfi/index.php/en/

The National Digital Library

The National Digital Library's Digital Preservation Solution

- A national, centralized, long term preservation solution of electronic materials
  - Scaleable as the volume and types of data, and number of partner organisations increase
  - Accommodates the increased volume and diversification of digital information and organisations
  - Provided and maintained by the state-owned CSC – IT Center for Science
- Built together with Digital Preservation Solution for Research Data

…But still: Why a centralised LTP solution?

- More cost efficient than several smaller systems
- Faster and more effective implementation for the partner organisations
- Enables seamless cooperation and shared usage across organisational boundaries
- Uniform standards and processes based on the best practices
- Guarantees the high quality of preservation actions, independently of the level of knowledge in participating organisations.
- Secure preservation, geographically decentralised to minimise threats

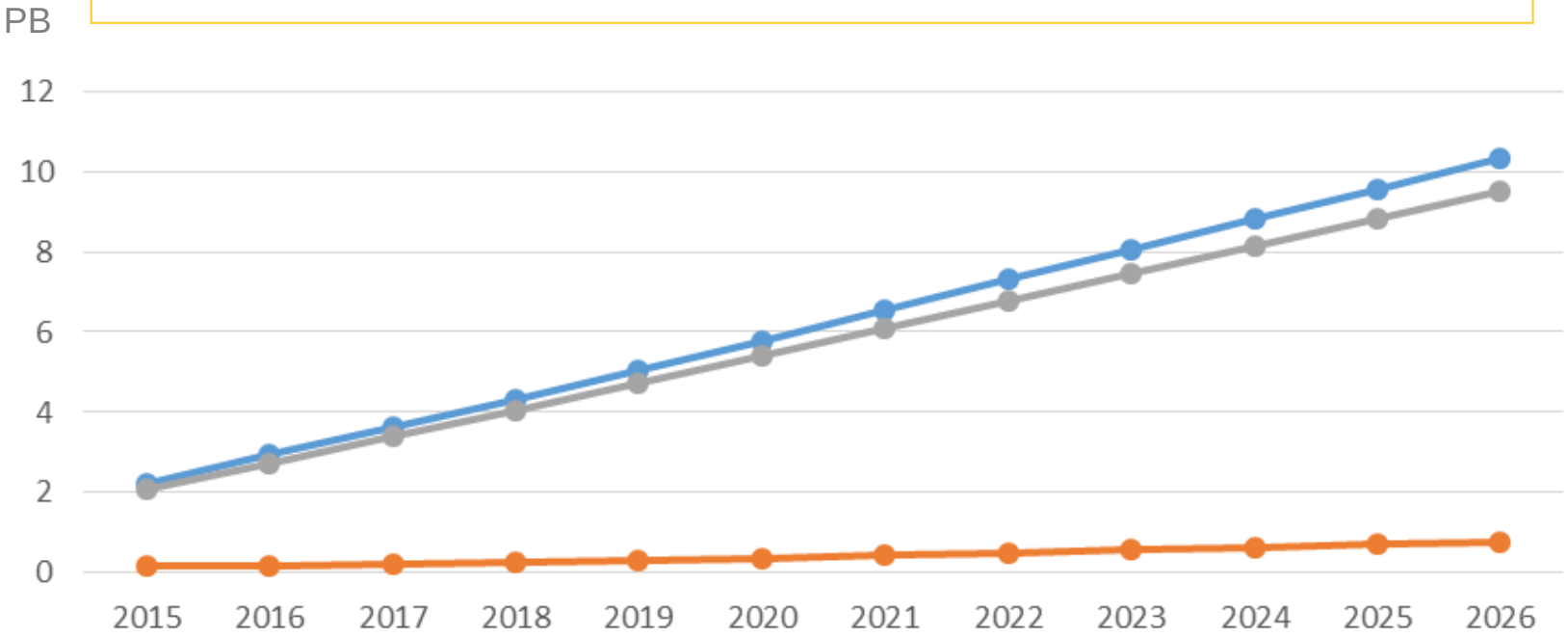
Digital Preservation Implementation Plan (2012)

Tradition of centralised services…

+ Shared benefit = total benefit
+ Biggest benefit for smaller organisations
- Not so flexible as doing things yourself

Estimation of digital content to be preserved

	2015 (PB)	2019 (PB)	2024 (PB)
Digitised content	2,07	4,73	8,15
Born-digital content *(Finnish Web Archive)*	0,16 *(0,08)*	0,31 *(0,14)*	0,65 *(0,24)*
TOTAL	2,23	5,04	8,80



Source: CSC (2015)

Collaboration

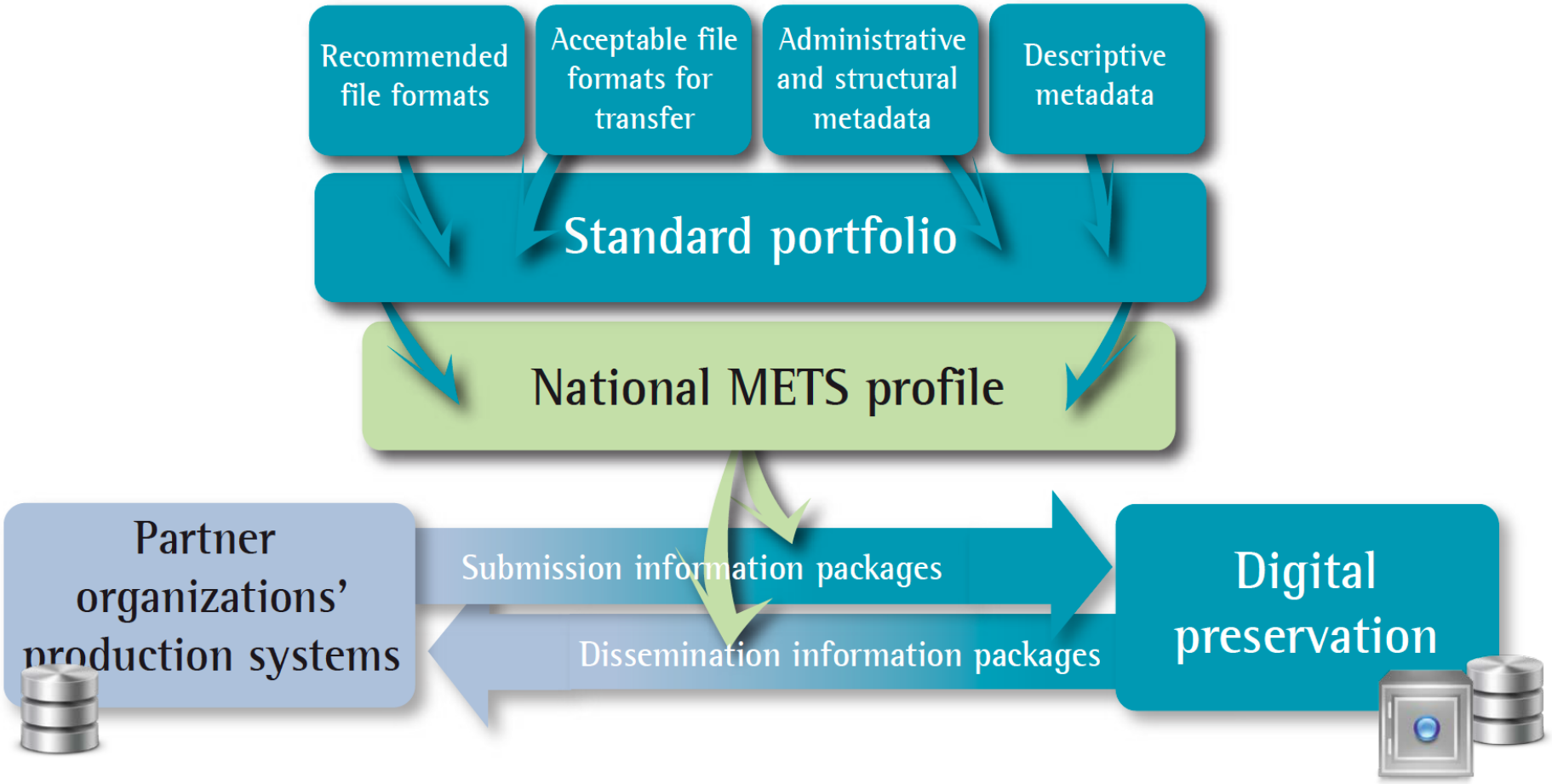
The National Digital Library:

- Steering Group, Enterprise Architecture Group, Metadata Branch…

NDL's Digital Preservation Solution Cooperation Group

- Participating in the development of the LTP solution and its functionality
- Preparing and updating the determination and profiles
- Promoting interoperability
- User perspective and experience
- Information sharing

Determination and profiles



Main Documentation for the Users of LTP Service

NDL Metadata Requirements and Preparing Content For Digital Preservation

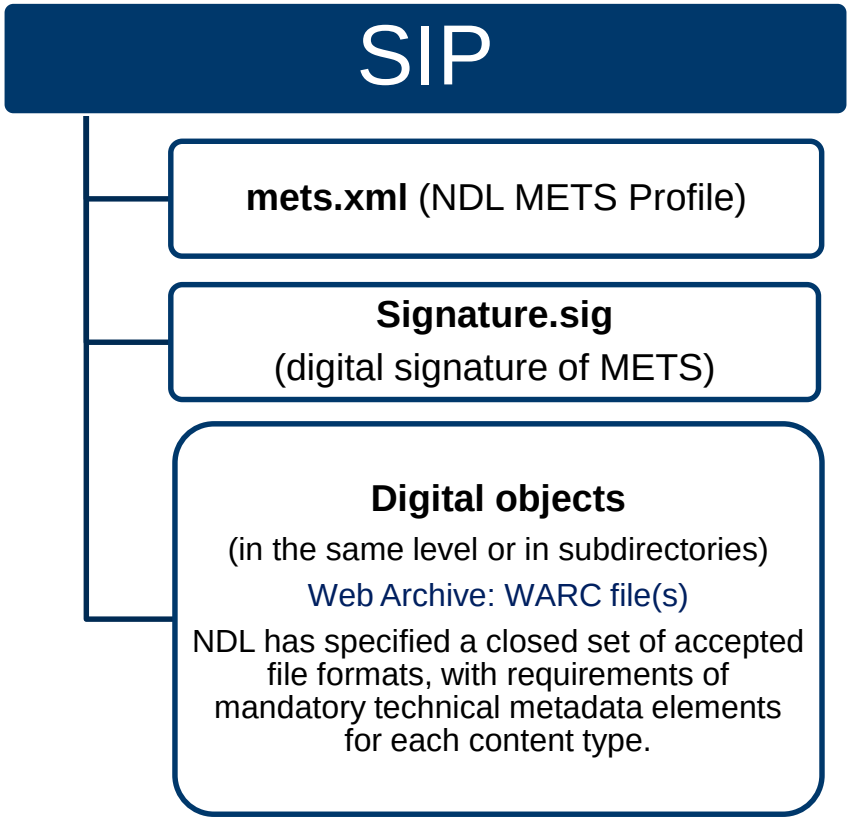
- Content Structure: NDL METS Profile
- Tecnical Structure of SIP
  - Content of SIP and Digital Signature
- Technical Structure of DIP

Acceptable File Formats for Preservation and Transfer

- Used as a recommendation already when publishing / digitizing materials

Interfaces of LTP Service

- … for transfering SIPs and receiving DIPs

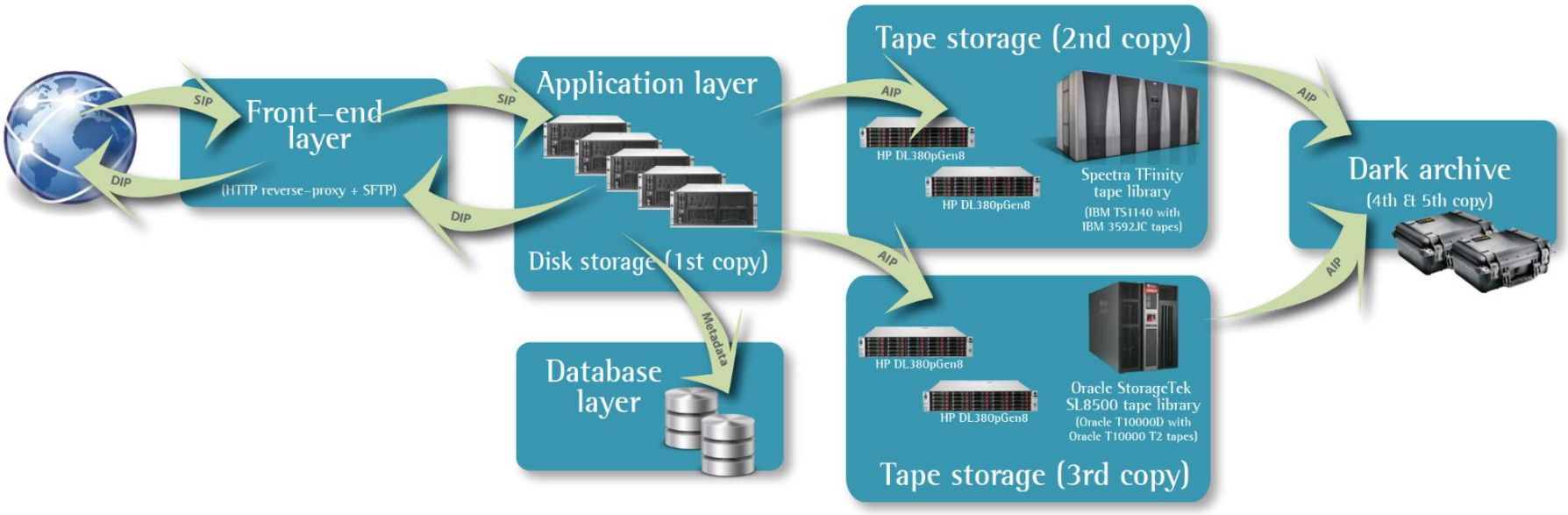


Acceptable File Formats for Preservation

(updated Jan 2016)

Text	• Comma Separated Values (CSV) • Electronic Publications (EPUB) • Extensible Hypertext Markup Language (XHTML) • Extensible Markup Language (XML) • Hypertext Markup Language (HTML) • Open Document Format (ODF) • PDF for long-term preservation: PDF-Archive (PDF/A) • Text file (Plain text)
Sound	• Audio Interchange File Format (AIFF), PCM-encoded • Broadcast Wave Format (BWF) • Free Lossless Audio Codec (FLAC) • MPEG-4 AAC – Advanced Audio Coding (AAC) • Waveform Audio Format (WAV)
Moving Picture	• JPEG 2000 sequence • MPEG-4 AVC – Advanced Video Coding (AVC)
Picture	• Digital Negative (DNG) • Joint Photographic Experts Group (JPEG) • Joint Photographic Experts Group JPEG 2000 (JP2) • Portable network graphics (PNG) • Tagged Image File Format (TIFF)
Web Archive	• **Web ARChive Format (WARC)**

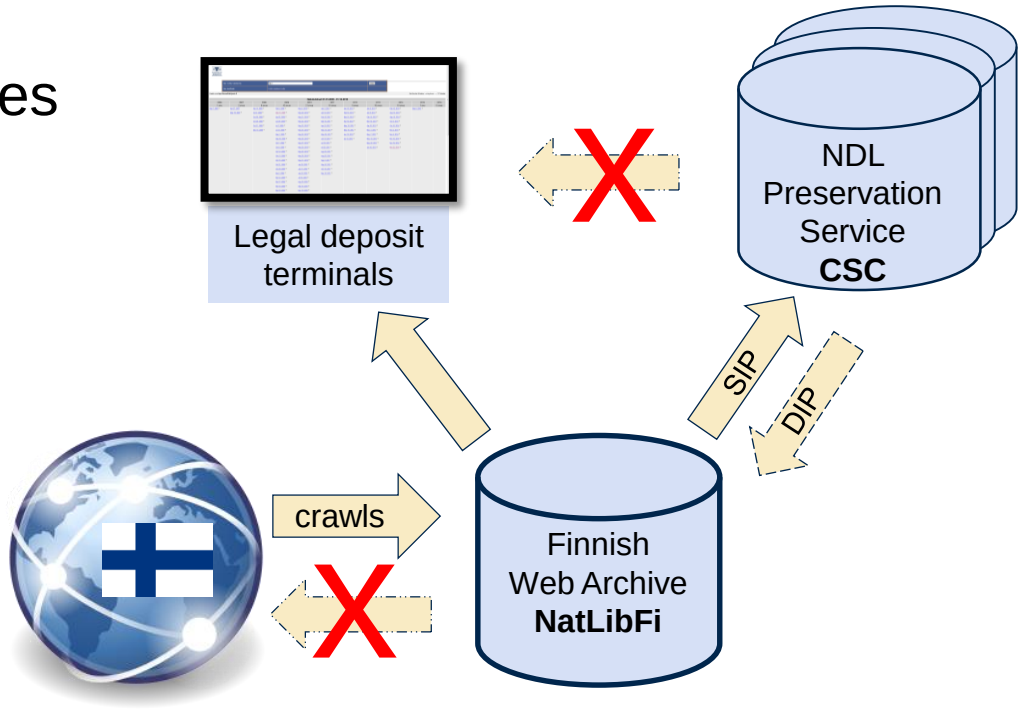
Preservation infrastructure



- Highly modular microservice structure
- Architecture is divided into small, highly independent 3rd party or in-house open source components.

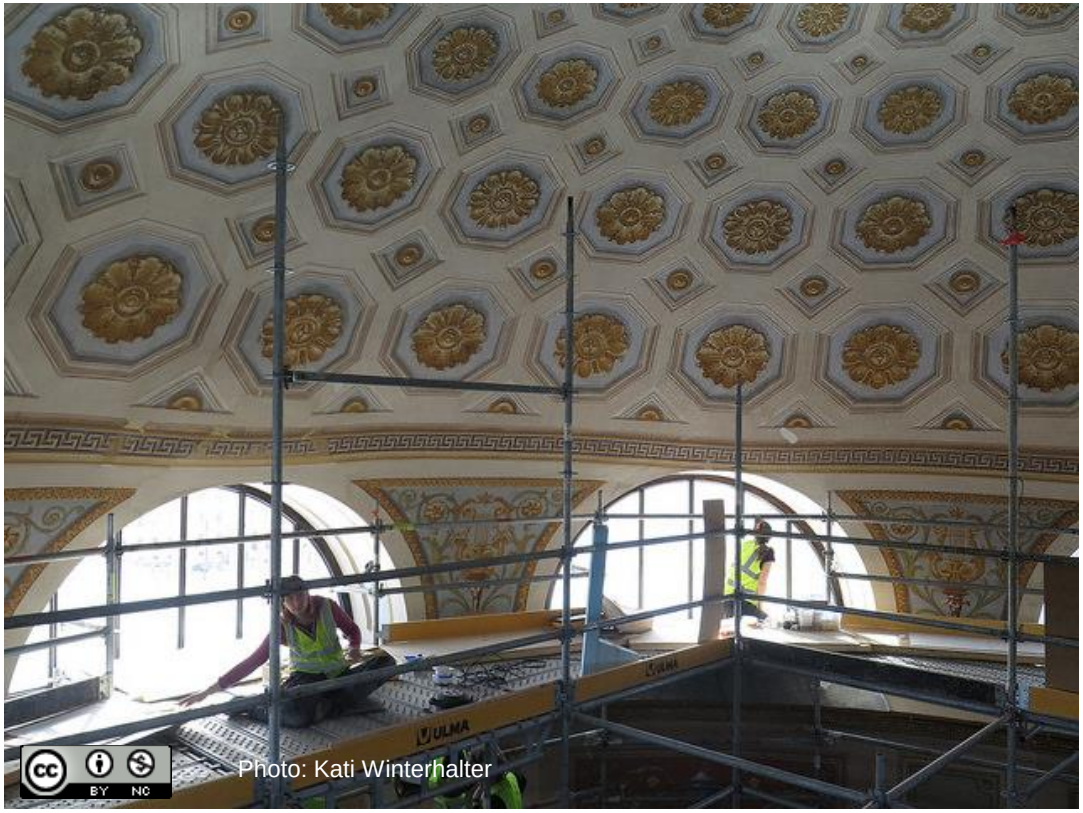
Finnish Web Archive and the Preservation Service

- Legal deposit terminals with NO other connections, USB ports etc
- Restricted use of interfaces
- Only authenticated users can create and download DIPs



Workflow of sending (Web Archive) SIPs to digital preservation

1. National Library creates SIP directories according to NDL specifications…
2. … and transmits the SIPs via SFTP to a buffer of the digital preservation service.
3. The workflow manager (in CSC) can then find the transferred data from the buffer and start processing it.
4. The ingest workflow has several microservices for validating the data according to the NDL specifications. Each of these microservices generate a validation report, which are finally combined as a final ingest PREMIS report.



Digital objects we send

- Going on:
  - (w)arc files
  - Logs and metadata about the Heritrix crawls
- To do:
  - Twitter JSON
  - YouTube JSON and video

The Heritrix crawl SIPs we send

- Warc files
  - directory for each warc file
  - each dir contains
    - mets.xml
    - warc-0001.warc.gz
    - signature.sig
- For the crawl metadata and logs
  - Again a new directory
  - Which has
    - mets.xml
    - list-of-warcs.xml
    - log-file.xml
    - signature.sig

Creating the SIPs - Warcs

- python script
  - lurks metadata
  - variables to add metadata
  - Creates the mets.xml files
- shell script for validation
- shell script for the signatures

Creating the SIPs – Crawl summary

- copypaste log files into a directory
- python script to create log.xml
- python script to create list-of-warcs.xml
- xml validation
- python script to create mets.xml
- shell script for the signature

… and the things to do

- Twitter JSON
  - text files are nice :)
  - more scripting

- YouTube videos and JSON
  - ^ for JSON
  - Flash videos should be converted
  - FFmpeg?
  - more scripting

Sending of the SIPs

- gzip and send each directory created
- sftp with rsa key
- SIPs go into the trasfer directory
- NDL-LTP service validation
- SIPs with validation reports end up
  - accepted directory
  - rejected directory
    - SIPs can be modified
    - Move into the transfer directory if everything is ok!

Retrieving of the DIPs

- REST API over HTTP
- TSL/SSL and HTTP Basic Access
- Search API supports searching by name or date, limit
- Request to retrieve a DIP by its ID
- Generate DIP
- Retrieve DIP when ready
- DIPs stored for 3 days in the disseminated directory

- We haven’t tried this yet ☺

NDL METS Profile (v.1.5)

Identifiers	WARC examples (application of METS profile for WARC)
Package Identifier (MUST)	<mets:mets OBJID="KK-20141001111035-00000-kerays-Suomi-FI-2014-Kansalliskirjasto">
Object Identifier (MUST)	<premis:objectIdentifierValue>KK-20141001111035-00000 …
Identifier for metadata (MAY)	- *{METS PID attribute + PIDTYPE}*
Timestamps	
Creation or Modification time of the Information Package (MUST)	<mets:metsHdr CREATEDATE="2015-11-23T14:23:27">
Creation Time of Digital Objects and Metadata Records (MUST)	OBJECTS: <premis:dateCreatedByApplication>2014-10-01… METADATA: <mets:dmdSec ID="dmdSec" CREATED="2015-11-23T14:23:27"> Descriptive Metadata Formats <mets:mdWrap MDTYPE="DC">

NDL METS Profile (v.1.5)

Technical metadata	WARC examples
File format (MUST)	<premis:formatName>application/warc</premis:formatName>
Version (SHOULD)	<premis:formatVersion>1.0</premis:formatVersion>
Fixity Information and its Algorithm (MUST)	<premis:fixity> <premis:messageDigestAlgorithm>MD5 </premis:messageDigestAlgorithm> <premis:messageDigest>2295bc529df58fe473bc768b5d7b1c99 </premis:messageDigest> </premis:fixity>
Technical characteristics (content type specific technical characteristics) (MUST)	{Heritrix version etc}

Restrictions	
Access restrictions, preservations restrictions (OPTIONAL)	-

NDL METS Profile (v.1.5)

Provenance information & structMap	WARC examples
Provenance information PREMIS event (MUST) PREMIS agent	<premis:event> <premis:eventIdentifier> <premis:eventIdentifierType>local</premis:eventIdentifierType> <premis:eventIdentifierValue>KK-2014100111103500000 </premis:eventIdentifierValue> </premis:eventIdentifier> <premis:eventType>creation</premis:eventType> <premis:eventDateTime>2014-10-01</premis:eventDateTime> <premis:eventDetail>{description about the crawl and its relations} </premis:eventDetail> <premis:eventOutcomeInformation> <premis:eventOutcome>success</premis:eventOutcome> </premis:eventOutcomeInformation> </premis:event>
Structural map (MUST)	<mets:structMap> <mets:div TYPE=”warc" DMDID="dmdSec"><mets:fptr FILEID=”warcfile"/> </mets:div> </mets:structMap>

Effects

- DIPs instead of tape backups
- Retrieving DIPs
  - Slow(?)
- No more tapes
  - Should we still keep our backups?
  - Less things to do and worry about?
- Everything stored it the NDL-LTP service is in good safe
- Crawl catalog, metadata and descriptions

How do You take care of the long term preservation of Your web archive?

- Please, let us know.

- You may also give us your opinion about the NDL-LTP service.

Thank You!

Lassi.Lager@helsinki.fi
Petteri.Veikkolainen@helsinki.fi

LIBER HELSINKI 29.6.–1.7.2016